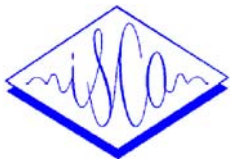

Speech Prosody 2004
Nara, Japan
March 23-26, 2004

# Perceptual Discrimination of Prosodic Types

*Masahiko Komatsu*[1], *Takayuki Arai*[1], *& Tsutomu Sugawara*[2]

[1]Department of Electrical and Electronics Engineering
[2]Faculty of Foreign Studies
Sophia University, Tokyo, Japan
koma2@splab.ee.sophia.ac.jp, {arai, sugawara}@sophia.ac.jp

## Abstract

A perceptual discrimination test was conducted to investigate whether humans can discriminate prosodic types solely based on suprasegmental acoustic cues. Excerpts from Chinese, English, Spanish, and Japanese, differing in lexical accent types and rhythm types, were used. From these excerpts, "source" signals of the source-filter model, differing in F0, intensity, and HNR, were created and used in a perceptual experiment. In general, the results indicated that humans can discriminate these prosodic types and that the discrimination is easier if more acoustic information is available. Further, the results showed that languages with similar rhythm types are difficult to discriminate (i.e., Chinese-English, English-Spanish, and Spanish-Japanese). However, detailed investigation of the results suggested the need for reconsideration of prosodic types from an acoustic and perceptual basis.

## 1. Introduction

It is known that humans can discriminate languages based on prosodic cues to some extent. A number of perceptual experiments have been conducted to investigate whether humans can identify or discriminate languages or dialects by hearing real speech sounds or processed/synthesized sounds that simulate the prosody of speech (see [1] for stimulus types). Although these experiments suggest that prosody plays a role in language discrimination, they have been conducted rather sporadically. It is not clear yet whether humans can perceptually discriminate various prosodic types, such as lexical accent types and rhythm types that linguists have referred to, and how they are related to the acoustic properties of speech. To investigate these questions, it is necessary to conduct perceptual experiments with the acoustic cues parameterized.

The present study investigated whether humans can discriminate lexical accent types (tone, pitch, and stress accents) and rhythm types (stress-, syllable-, and mora-timed) and how they are related to acoustic properties.

In the perceptual experiments, we used the "source" of the source-filter model as the stimuli. In their synthesis process, we controlled F0, intensity, and Harmonics-to-Noise Ratio (HNR). The temporal patterns of F0 and intensity are undoubtedly related to prosody. Besides, the source is related to the sonority feature (broad classification of phonemes) that seems to be an important contributor to rhythm. Recent studies on rhythm, such as [2-4], are based on the durations of consonant and vowel intervals, which means the acoustic properties that discriminate such phoneme classes are relevant. Ramus et al. ([3], p. 271 fn) wrote "[their] hypothesis should ultimately be formulated in more general terms, e.g. in terms of highs and lows in a universal sonority curve" rather than the consonant-vowel distinction. The source of the source-filter model significantly contributes to the perception of the sonority feature [1, 5].

## 2. Speech data

We chose Chinese, English, Japanese, and Spanish as the languages representing prosodic types. Fig. 1 shows a provisional schematic layout of their lexical accent and rhythm types. In the figure, Chinese is situated at "tone accent" and "stress rhythm" tentatively because it is said to have both lexical tones and stress [6] although it is not traditionally classified in terms of rhythm [4].

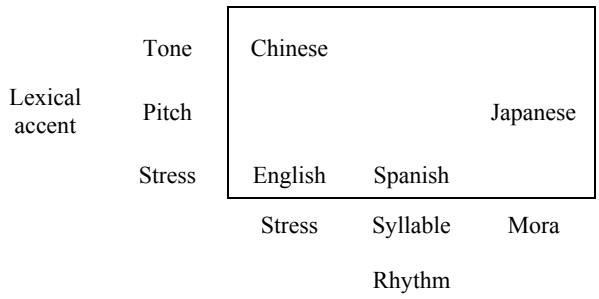


Figure 1: *Provisional schematic layout of prosodic types.*

The speech samples of English and Spanish were drawn from the MULTEXT prosodic database [7, 8]. The corpus consists of the recordings of 40 different passages for each of English, French, Italian, German, and Spanish. The same passages are translated into each language. The translation is rather free, and the expressions such as proper names are often changed to be adapted to the local culture. Besides the speech recording, the corpus contains the data of the stylized F0 curves created by the MOMEL algorithm [9], which extracts the macroprosodic component from the original F0.

Japanese samples were taken from the Japanese MULTEXT β [10, 11]. It has the "play" and "read" versions, the latter of which was used in our experiment. This corpus also contains the MOMEL curves, but the data for one speaker were missing in the corpus, and we created substitutes for these.

We recorded the Chinese samples ourselves. The passages were translated into Mandarin by a Chinese collaborator. The recording was conducted in a soundproof studio (microphone: Sony ECM-MS957; DAT recorder Sony TCD-D100). Later, the audio data were down-sampled to 22 kHz (Onkyo SE-U77). When the F0 contours were stylized, the MOMEL algorithm was slightly modified (see Section 6).

In the present experiment, for each language, 9 passages read by 3 speakers were used (Passage IDs *o6-o8* by Speaker

10.21437/SpeechProsody.2004-168

1, *p1-p3* by Speaker 2, and *p6-p8* by Speaker 3). The same passages were used across the languages to avoid emotional or attitudinal differences. Passages read by female speakers were selected because their pitch range is wider than that of males, and we expected the prosodic differences to be more distinct. The experimental stimuli were made from the first 5 seconds of these selected passages.

## 3. Experimental Procedure

### 3.1. Signal processing

Six types of stimuli were created from the original speech. They simulated some characteristics of the original speech as described in Table 1. These sets can be grouped into three: those carrying amplitude information (Sets 1-3), the one carrying pitch information (Set 4), and those carrying both of them (Sets 5-6).

Table 1: *Stimulus sets. The mid column indicates what the stimuli simulate, and the right column indicates what they are made of.*

| | simulates | is made of |
|---|---|---|
| Set 1 | Intensity | white noise |
| Set 2 | Intensity | pulse train |
| Set 3 | Intensity, HNR | white noise and pulse train |
| Set 4 | F0 | pulse train |
| Set 5 | Intensity, F0 | pulse train |
| Set 6 | Intensity, HNR, F0 | white noise and pulse train |

Set 1 is made of white noise.

Set 2 is made from a pulse train, whose F0 was constantly set to the mean value of the MOMEL curve.

Set 3 is a mixture of white noise and a pulse train. The amplitude contours of a harmonics component and a noise component of the original signal were calculated respectively. Then a pulse train was made based on the amplitude contour of harmonics, white noise was made based on the amplitude contour of noise, and they were added together. F0 of the pulse train was constant as well as Set 2. In sum, voiced intervals in the original signal were represented as close to pulse trains; and unvoiced intervals, as close to white noise.

Set 4 is the pulse train created from the MOMEL curve. All unvoiced intervals were interpolated by MOMEL. The intensity was set constant.

Set 5 is the same as Set 4 except that it simulated the intensity of the original signal.

Set 6 is the mixture of white noise and a pulse train. It is the same as Set 3 except that it carried the stylized F0 contour. Note that unvoiced intervals such as [s] in the original signal did not carry F0 in Set 6 because they were converted to white noise, while they did carry interpolated F0 in Sets 4 and 5 because the whole signal was made of a pulse train.

All of these sounds were created with Praat (Version 4.1.6). They were created at the sampling frequency of 16 kHz. If amplitude less than a certain threshold continued more than 200 ms in the original signal, such intervals were regarded as pauses and suppressed to silence in the synthesized signal. Finally, they were tilted by −6 dB/oct to make them sound like a human voice.

Spanish data carried an unfavorable noise (seemingly a hum noise). Before the processing above, a noise reduction process (CoolEdit 2000) was applied only to the Spanish data.

We used the MOMEL data in the corpus if it existed. We calculated the MOMEL curve for one passage of Japanese which was accidentally missing from the corpus. We also calculated such curves for the Chinese speech. The Praat script (by G. Rolland, 2000; modified by S. Werner, 2002) was used for calculation, but the modification was necessary for Chinese (see Section 6).

### 3.2. Perceptual experiment

10 graduate students and researchers (age: 23-39) specializing in linguistics, speech therapy, or speech engineering, voluntarily participated in the experiment. We asked those who were experienced in listening to various speech sounds because it was expected that the task would have been difficult for non-specialists.

The experiment was conducted in a soundproof studio. Stimuli were provided from a personal computer through a digital audio processor (Onkyo SE-U77) and headphones (Audio-Technica TH-65). Participants were allowed to adjust the volume of the stimuli according to their taste. The experiment was done with Praat.

Before the test session, sample sounds were given to each participant for a demonstration. The samples were, for each language, one original sound and Set 1-6 sounds created from it. The participant was asked to listen to all original sounds and at least one from each of Set 1-6 sounds. Then, the participant went through a short training session to become familiar with the operation of the program.

In the test session, the participant was asked to listen to a language pair and judge in which order the languages were presented. For example, after clicking the mouse, the Set 1 sound of Chinese and the Set 1 sound of English were successively played, and the participant clicked one of the two alternatives on the screen, namely, "(1) Chinese – (2) English" and "(1) English – (2) Chinese". The pairs were made so that they have the same passage ID, e.g. Chinese *o6* and English *o6*.

The test session was conducted from Set 1 to Set 6. Each set consisted of 6 subsets: Chi-Eng, Jpn-Spa Chi-Jpn, Eng-Spa, Chi-Spa, and Eng-Jpn. Each subset consisted of 6 trials (3 passage pairs [3 passages by different speakers, e.g. *o6*, *p1*, and *p6*] × 2 presentation orders [e.g. Chinese-English and English-Chinese]). The order of subsets within a set and the order of trials within a subset were changed for each participant. In total, the test session had 216 trials (6 trials × 6 subsets × 6 sets) and continued for about 1 hour.

## 4. Results and discussion

### 4.1. Analysis by Multi-Dimensional Scaling (MDS)

The rates of correct responses were regarded as the distances of the stimulus pairs. If the pair was identified well, the stimuli were discriminated well. However, it does not make sense to say the correct response rate of 50% (chance level) implies a better perceptual discrimination than the lower rates. Therefore, the raw rates were subtracted by 50%, and regarded as 0% if they are negative. That is, the chance level was converted to the distance of 0, and the rates lower than chance were also converted to 0.

These converted values were input to the MDS procedure of SPSS Version 11.0.1 (ordinal data, symmetric, untie tied observation, Euclidean distance model, 2-dimensional). The

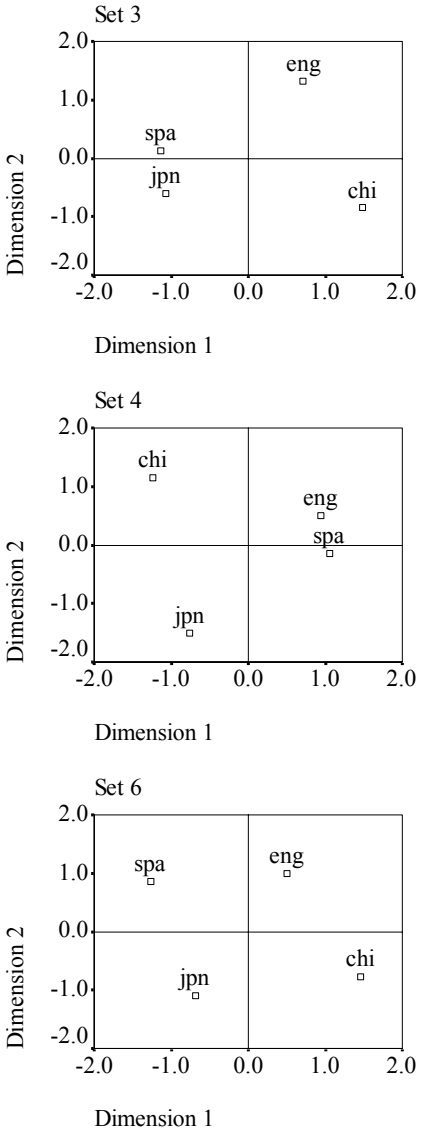


Figure 2: *Derived configurations of perceptual distances. Calculated from the results of Set 3 (top), Set 4 (middle), and Set 6 (bottom).*

configurations constructed from the results of Sets 3, 4, and 6 are shown in Fig. 2 (Young's S-stress formula 1 = .154, .141, .097; Kruskal's stress formula 1 = .112, .109, .052; and $R^2$ = .860, .880, .931; respectively).

The top panel shows the results for Set 3 stimuli, which have only the amplitude information. It indicates that such information does not suffice to discriminate Spanish (syllable-timed) and Japanese (mora-timed), whose rhythm types may be regarded as similar.

The middle panel shows the results for Set 4, which has only the F0 information. English and Spanish, both of which are categorized as having stress accent, are located close to each other.

The bottom panel shows the results for Set 6, which has both amplitude and F0 information. All four languages are discriminated from one another.

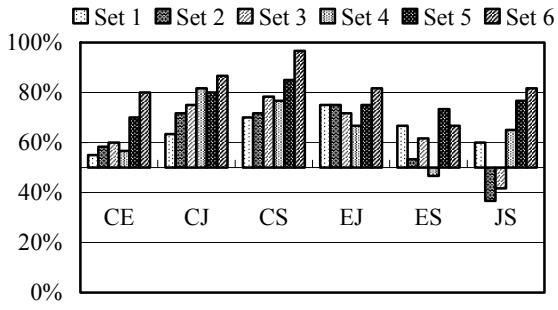


Figure 3: *Correct response rates for each language pair. "CE" stands for Chinese-English, etc.*

### 4.2. Analysis by individual correct response rates

In general, as the available information increases, the rates of correct responses increase, that is, discrimination gets easier. Correct response rates averaged across all languages are 65.0%, 61.1%, 64.7%, 65.6%, 76.7%, and 82.2% from Sets 1 to 6. The rates for Sets 1-3, those with only amplitude information, are comparatively low, and the rate for Set 4, the one with only F0 information, is also low, but when such information combines (Sets 5-6) the rates get higher. This result is quite straightforward.

Looking into the correct response rates of each language pair, the situations get complicated. Fig. 3 shows the correct response rates for each language pair.

The Chinese-Japanese pair (CJ) and the Chinese-Spanish pair (CS) have good scores; they are easy to discriminate. This is understandable because these pairs are different both in lexical accent type and rhythm type. In the Chinese-Japanese pair, F0 (Set 4) seems more effective than amplitude (Sets 1-3), but this is not the case with the Chinese-Spanish pair. This may be interpreted to mean that F0 is a cue to discriminate stress and non-stress languages. (The word "language" is used here because it is not clear whether this classification should be attributed to lexical accent types or rhythm types.)

The English-Japanese pair (EJ) has comparatively good scores, but F0 does not have much contribution in this pair. Set 4 (F0 alone) is lower than Sets 1-3, and Set 5-6 does not show much improvement compared to Sets 1 and 3 (Their difference is the presence or absence of F0). This may be interpreted to mean that F0 is not a strong cue in the discrimination of stress and pitch accents.

The Chinese-English pair (CE) is difficult to discriminate. It is noteworthy that F0 (Set 4) failed to discriminate tonal and non-tonal accents. The difficulty probably derives from the fact that both languages have stress accent or rhythm. (Remember F0 worked to discriminate stress and non-stress languages with the Chinese-Japanese pair.)

The English-Spanish (ES) and Japanese-Spanish (JS) pairs are difficult to discriminate. Sometimes the rates go down below the chance level, which may suggest that the listeners were quite confused with these stimulus pairs. Considering the rhythm continuum [3], it is reasonable to suggest that English-Spanish and Japanese-Spanish are more difficult to discriminate than English-Japanese.

Considering all these together, rhythm types seem to be important to determine the discrimination difficulty. Even F0 seems to contribute to this by sometimes discriminating stress

and non-stress languages. It seems that F0 alone cannot be a discriminator of lexical accent types.

Finally, consider the effects of HNR. The difference between Sets 1 and 3 and the difference between Sets 5 and 6 are the presence or absence of HNR information. While Set 3 shows higher rates than Set 1 only in 3 language pairs (CE, CJ, CS), Set 6 shows higher rates in 5 language pairs (all except ES). That is, the same information brought more improvement when combined with F0. It is inferred that it was easier in Set 6 to capture the timing relation of F0 change with the occurrence of some units such as syllables (e.g., whether or not an F0 change is within a syllable) than in Set 5 (and also Set 4) where F0 were interpolated during unvoiced intervals.

## 5. Conclusions

MDS procedures and averaged correct identification rates produced results that conform to a rather common-sense view. Humans can discriminate lexical accent types and rhythm types. The more acoustic cues that are available, the easier discrimination is. The lack of F0 information makes the discrimination of Spanish and Japanese, languages with similar rhythm types, difficult. Likewise, the lack of amplitude information makes the discrimination of English and Spanish, which have the same lexical accent types and only differ in rhythm types, difficult.

The individual inspection of language pairs showed that the discriminations of Chinese-English, English-Spanish, and Spanish-Japanese are difficult, a finding that accords with the linguistic categorization of prosodic types.

However, further investigation raised questions. The results suggested that rhythm types are more easily discriminated than lexical accent types and that even F0 contributes to the rhythm discrimination. In such circumstances, we do not yet know whether the distinction of stress and non-stress languages should be handled as the opposition of rhythm types or that of lexical accent types or both, and further how independent from or dependent on each other the rhythm types and lexical accent types are. We need to reconsider the layout of prosodic types in Fig. 1 from an acoustic and perceptual basis.

## 6. Appendix: Applying MOMEL to a tone and a pitch accent language

The MOMEL algorithm is designed to remove microprosody, but this seems to cause inconveniences if it is used to resynthesize tone and pitch accent languages. It often removes F0 movements necessary for lexical accents.

The algorithm approximates the F0 contour by a quadratic regression within an analysis window, which is typically 300 ms [9]. This window length seems too long for typical Chinese utterance, which has quicker up-and-downs than non-tonal languages. This setting often removes lexical tones, leaving only phrasal intonation. To capture lexical tones, we shortened the analysis window length to 100 ms. However, if we set the length so short, it seems that the target points too easily fall outside the analysis window and subsequently are neglected. Especially, phrase initial steep F0 falls tend to be neglected. Although the search region of the target points is defined as the same as the analysis window in [9], we extended it to 4 times of the analysis window, i.e. 400 ms, and it produced better results. In the present study, we set the analysis window to 100 ms, the search region of targets to 400 ms, and the reduction window (typically 200 ms) to 100 ms.

In applying the algorithm to Japanese, it seems that the typical setting often distorts the lexical accent, especially phrase initially. However, we used the typical setting for our analysis of the passage of Japanese because we did not want to process it in a different manner than the MOMEL data contained in the corpus.

## 7. References


[1] Komatsu, M., 2002. What constitutes acoustic evidence of prosody? The use of LPC residual signal in perceptual language identification. In *LACUS Forum 28,* Brend, R.M.; Sullivan, W.J.; Lommel, A.R., eds. Houston, TX: Linguistic Association of Canada and the United States, 277-286.

[2] Ramus, F.; Mehler, J., 1999. Language identification with suprasegmental cues: A study based on speech resynthesis. *Journal of the Acoustical Society of America,* 105(1), 512-521.

[3] Ramus, F.; Nespor, M.; Mehler, J., 1999. Correlates of linguistic rhythm in the speech signal. *Cognition,* 73(3), 265-292.

[4] Grabe, E.; Low, E. L., 2002. Durational variability in speech and the Rhythm Class Hypothesis. In *Laboratory Phonology 7,* Gussenhoven, C.; Warner, N., eds. Berlin: Mouton de Gruyter, 515-546.

[5] Komatsu, M.; Tokuma, S.; Tokuma, W.; Arai, T., 2002. Multi-dimensional analysis of sonority: Perception, acoustics, and phonology. In *Proceedings of International Conference on Spoken Language Processing 2002, Denver, CO,* 2293-2296.

[6] Hirst, D.; Di Cristo, A., 1998. A survey of intonation systems. In *Intonation Systems: A Survey of Twenty Languages,* Hirst, D.; Di Cristo, A., eds. Cambridge, UK: Cambridge University Press, 1-44.

[7] Campione, E.; Véronis, J., 1998. A multilingual prosodic database. In *Proceedings of International Conference on Spoken Language Processing '98, Sidney, Australia,* 3163-3166.

[8] Campione, E., 1998, ed. *MULTEXT prosodic database* [CD-ROM]. Paris: European Language Resources Association.

[9] Hirst, D.; Di Cristo, A.; Espesser, R., 2000. Levels of representation and levels of analysis for the description of intonation systems. In *Prosody: Theory and Experiment,* Horne, M., ed. Dordrecht, The Netherlands: Kluwer Academic, 51-87.

[10] Kitazawa, S.; Kitamura, T.; Itoh, T., 2002. Nihongo MULTEXT ni okeru inritsu joho no bunseki to shuroku. In *Proceedings of 2001 2nd Plenary Meeting and Symposium on Prosody and Speech Processing, Tokyo,* 39-50.

[11] Kitazawa, S., 2002, ed. *Japanese MULTEXT* (β version) [CD-ROM]. Shizuoka University, Japan.



* We thank Yin WenYi, Kanae Amino, and Makiko Aoyagi for their help in experiments, and Terri Lander for her comments. We also thank Hélène Loevenbruck and Stefan Werner for the information on the Praat script.